

DOCUMENT RESUME

ED 309 649

FL 018 121

AUTHOR Carroll, David; Kowitz, Johanna
TITLE Text Concordancing--What It Is and What It Can Do for You.
PUB DATE 2 Apr 89
NOTE 17p.; Paper presented at the Annual Meeting of the International Association of Teachers of English as a Foreign Language (23rd, Warwick, England, April 1989).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Techniques; *Computer Software; Computer Uses in Education; *Discourse Analysis; Instructional Materials; Language Tests; Lexicography; Material Development; Morphemes; Search Strategies; Second Language Instruction; Textbook Content; *Word Processing
IDENTIFIERS *Concordance (Data)

ABSTRACT

A text concordancer is a computer program that operates on a body of text to locate and sort all occurrences of a target morpheme, word, or pattern. Developments in word processing and hard disks in small computers have made new applications of the technique possible with direct relevance to the classroom. A discussion of the technique looks at the properties of a good concordancing program and briefly surveys some available programs. Features discussed include: the ability to accept data in many different word processing formats and ASCII; flexibility in selecting texts or sections of a text for analysis, using "tags"; range of search strategies available, including searching for phrases, and use of "wild cards"; and availability of different types of contextualization. The study discusses possible applications to the development of materials, tests, and classroom examples and illustrates these applications with a variety of corpora. The bibliography includes 33 items. (Author/MSE)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED309649

TEXT CONCORDANCING - WHAT IT IS AND WHAT IT CAN DO FOR YOU

David Carroll, NCERD, Cairo, Egypt
Johanna Kowitz, USIA, Washington DC

Paper presented at the Annual Meeting of
the International Association of Teachers
of English as a Foreign Language (23rd,
Warwick, England, April 1989).

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to improve
reproduction quality.
- Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Carroll, D.

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

FL018121

TEXT CONCORDANCING - WHAT IT IS AND WHAT IT CAN DO FOR YOU

David Carroll, NCERD, Cairo, Egypt
Johanna Kowitz USIA, Washington DC

ABSTRACT

A concordancer is a program that operates on a body of text to locate and sort all occurrences of a target morpheme, word or pattern. Developments in word processing and hard disks in small computers have made possible new applications of this technique with direct relevance to the classroom. This paper considers the properties of a good concordancing program, with a brief survey of some available programs. Features discussed include: ability to accept data in many different word processor formats as well as ASCII; flexibility in selecting texts, or sections of a text for analysis (using "tags"); range of search strategies available, including searching for phrases, and use of "wild cards"; availability of different types of contextualization - KWIC context, sentence/paragraph; sorting - left, right and backward. Possible applications to the development of materials, tests, and classroom examples are discussed and illustrated with actual corpora used by the authors. Examples include: large samples of real-life usage (e.g. Grolier's Electronic Encyclopedia); "home grown" corpora based on textbooks, teacher's books, transcripts of classroom tapes, student and academic writing samples.

A. INTRODUCTION

There are many new computer applications to language teaching and research. In most of them, however, the amount of help the computer can give to the language teacher or researcher is still very limited. For example, while grammatical analysis (and hence parsing) has a lot of potential for the future, at the moment it still requires grammatical tagging of samples, which severely limits the size of text that can be processed. Leech (1986) describes the process of tagging; Atwell and Elliott (1987) describe some of its potential uses.

On the other hand, lexically-based applications, involving locating and counting occurrences of words or patterns, are available now. Such applications work mainly on correct or corrected data (cf. Atwell and Elliott (1987) and Booth (1987) on the problems of data), to produce a concordance. A concordance, according to the Collins COBUILD dictionary, is "an alphabetical list of the words in a book or a set of books which also says

where each word can be found and often how it is used." It has three elements - a wordlist, an index and a set of contexts.

Concordancing is a well-established tool in literary analysis and Bible study. Such concordances have long been made manually; the texts involved, even if they are as long as the Bible, are limited in both number and length, and have a certain permanence. When a concordance of a literary or religious text has been made, it does not need to be repeated; so the effort involved is justified, in terms of the possibilities for text analysis.

B. SURVEYS OF USAGE

Students of literature work with texts of lasting significance; a concordance of a literary text is done once. Linguistic researchers, on the other hand, work with text that (at least for the researcher) is only a sample from a larger domain. Therefore linguistic research, to derive empirical findings about the occurrence and patterns of words, requires large samples of text to be gathered from diverse sources. These samples are called corpora (hence concordancing is also called "corpus browsing").

The principles for selection of material vary. Some corpora, such as the Collins-Birmingham University International Language Database (COBUILD), take complete texts; others, such as the Lancaster-Oslo/Bergen (LOB) or London-Lund, use samples of standard length (2,000 words for LOB; 5000 for London-Lund). Samples are taken from a variety of text-types in order to secure diversity and representativeness. Francis (1979) and Leech (1987) describe the composition of the Brown and LOB corpora, and also the difficulties encountered in assembling early corpora (this was especially true with the LOB). Svartvik and others (1982) describe the development of the London-Lund corpus.

Most corpora do not mix media - they are either spoken or written, - and usually contain one variety of English - British, American, Australian etc. Corpora of spoken English tend to be smaller than corpora of written English, because of the difficulty of transcription. At present, the typical overall size for a corpus of written English will be from one to five million words. The largest currently in use is the COBUILD, at around 20 million words (this is held on a mainframe computer).

Some corpora are available to researchers, either as "hard" (i.e., paper) copy, or on computer tape, diskette or microfiche. They can be obtained as raw data, or with grammatical tags (for research on grammar and parsing), or as complete concordances. The International Computer Archive of Modern English (ICAME) has a range of corpora which can be accessed.

Corpora currently available from ICAME include the Brown

(American written English), LOB (British written English), London-Lund (British spoken English) and Melbourne-Surrey (Australian newspapers). The contact point for these corpora is the Norwegian Computing Centre for the Humanities, in Bergen (one of the three ICAME institutions). The Oxford Text Archive also contains a collection of computer-readable files containing mis-spellings of English words, available to researchers. Mitton (1986) describes the content of this collection.

C. PROPERTIES OF A GOOD CONCORDANCER:

By the mid 1970's, programs had been developed for analyzing corpora and preparing frequency lists, frequency profiles, concordances, collocational analysis, and text analysis statistics. Sinclair (1986) describes these techniques and the development of software. A concordancing program can provide the user with four main types of output:

- a. **a wordlist:** with frequencies. This may be arranged either alphabetically or in order of frequency. It may also be grouped under "headwords" (lemmatization), in order to give a clearer indication of the total number of different words.
- b. **an index:** a wordlist with distributional information (frequencies and line references) but without any context. As with the wordlist, an index can be arranged either alphabetically or in order of frequency.
- c. **a concordance:** consisting of all occurrences of selected items, with the contexts in which they occur. For most linguistic applications, Key Word in Context (KWIC) format is used. This consists of the key word, usually aligned in a column, with context to either side - the depth of context being determined by the researcher. For some purposes, however, the context may instead be a natural unit (sentence or paragraph). The concordance can be extended to include analysis of collocation - that is, a list of all the cases where two specified words occur within a given interval (e.g., five words).
- d. **statistics:** a concordance program will normally give: a frequency profile giving cumulative total, relative frequencies, vocabulary and word totals; and type/token ratio. It does not normally give readability statistics (e.g. average sentence length, word length, number of syllables, proportions of different kinds of words, readability indices, etc.). These can be obtained from a specialist program (e.g. Kern International's Readley, or Buttonware's PC-Style).

D. APPLICATIONS OF CONCORDANCING TO ELT:

This paper suggests three main types of possible applications to ELT. First, there are those that improve the quality of published materials; second, those that arise from the power to process large bodies of published text; finally, there are applications in research, materials writing or lesson preparation involving processing of locally-gathered text.

1. Applications to published materials

With the exception of COBUILD, most available corpora have not been much exploited. For example, the work on the Brown Corpus carried out at Brown University (e.g., Francis and Kucera, 1982) has not led to much use of this corpus elsewhere. There are several possible reasons for this lack of interest. The first is the limited size of existing corpora. A million words might seem impressive; but for generalizing about lexical items, twenty million words is not excessive. Secondly, because in Britain and the U.S. linguistics has tended to be speculative rather than empirical, the possibilities of large-scale data processing have been rather neglected. This is less true of other parts of the world, notably Scandinavia, where both English and native languages have been studied statistically. Andersson (1978) and Lofstrom (1978) describe Swedish work in this area. Svartvik (1982) reports the activities of the survey of spoken English at the University of Lund (Sweden), and research projects in process. The third, and probably most important, reason for lack of interest is that, for much of the life of these corpora, the computing power required to process them has not been available to individual researchers. It is only in the past few years that the mass storage and processing speed required have been available on small computers. However, the possible applications are endless. Three main areas suggest themselves:

- a. **Lexicography:** Now that concordancing is possible on a desktop computer, interest is growing rapidly. The first benefits have become apparent in lexicography. Ideally a lexicographer would have access to unlimited amounts of data; but access has been limited by the need for humans to read, extract examples and make generalizations. As a result, the "dictionary definition" of a word has usually been based on the lexicographer's educated judgment. This is more likely to give the "root" meaning than show the way in which the item is actually used. Sinclair (1987 : p. xvii) gives examples of this. A second benefit to lexicography is that while certain kinds of information are relatively easy to gather - such as evidence about when usages were first seen, and about a word's derivation, other kinds of information - about the range of meanings a word may have; the structural patterns in which it occurs; about collocation; or about the relative frequency of use of

various meanings - have in the past been much harder to obtain. Gathering contemporary information of this kind was a lifetime's work, and hence out of date by the time it was published. Therefore, many dictionaries concentrate on etymology, historical information, and a priori definitions of words. Such information is useful; but not ideal for learners, who usually want to know what a word means, and how it is used.

Furthermore, dictionaries often use constructed examples of usage in support of their definitions. There are honorable exceptions to this, notably the Oxford English Dictionary, and Johnson's Dictionary. The disadvantage of constructed examples is that "they have no authority apart from the thoughts of the person who creates them and are often quite unnatural." (Sinclair, 1987). Such "examples" are not necessarily examples of good practice, although they may help the reader to understand the explanations. Controlled examples are especially unsuitable in dictionaries that are designed for language learners.

Computer concordancing has therefore made possible these three major advances in lexicography. First, information from collocation is easily accessible - information about the associations of a word and contexts in which it is found as well as its "signification". Second, statistical information is available to enable meanings to be ordered according to their frequency of occurrence. Third, examples can be drawn from genuine text to illustrate and amplify definitions. Sinclair (1985) describes the basic methodology of lexicographic observation of language in use, and the limitations on introspection and received description. He distinguishes citation of instances from concordancing of texts (in terms of statistical validity) and introduces the concepts of a corpus, lemmatization and collocational analysis. Benson (1985) describes collocation in greater detail. He compares collocation with idiom, showing how they differ; and draws attention to the need for more thorough analysis of collocation, and making the results available in dictionaries.

The Collins COBUILD Dictionary (Collins COBUILD, 1987) is the first dictionary to be based on empirical data of this kind, the Collins-Birmingham COBUILD corpus. It is radical in several respects. First, it gives extra attention to common words with many meanings. Second, it includes functional information. Third, it organizes entries according to frequency, as well as central or core meaning.

- b. **Specification of course/textbook content:** A second major possible application of concordancing techniques is specification of content, both linguistic and situational,

for course and teaching materials. A number of studies have considered the analysis of corpora as a tool for defining linguistic content of a course, especially for ESP/EAP. Khaled (1984) and Pickett (1986) both emphasize the value of empirical study of characteristic text in defining linguistic content; and Kennedy (1985, 1987) analyzes the Brown and LOB corpora to discover empirically how a communicative notion - "frequency of occurrence" - is expressed in English. Macdonald and others (1982) show how the computer analysis of job-relevant text materials can be used for developing materials and reducing time spent on lesson planning; and how these procedures can be used by persons with no prior computer experience.

The value to textbook writers of genuine contextual information cannot be overstated. Although the practice has been castigated by applied linguists for a number of years, many textbooks still depend largely on constructed examples. These are better for explaining meaning than teaching how language is used. Access to genuine data allows textbooks to be built around genuine, rather than constructed examples. This approach has been used in designing the Collins COBUILD English Course (Willis and Willis, 1987), a general English course using the COBUILD database.

- c. **Textbook evaluation:** The same techniques that can be applied to textbook writing can be used to monitor empirically the content of existing textbooks, as part of the selection or evaluation process. Given a computer-readable text, a concordancer can be used to list the words/structures that occur, and give their frequencies and contexts. This information can be used to assess whether or not: the syllabus is covered; important words and structures are given sufficient exposure; "productive" and "receptive" items are used in ways that reflect the distinction; contexts in which items are used reflect natural language (using the findings from corpus studies); the principles of a "spiral" syllabus are followed, in that items recur during the course of the book in different or more complex contexts. This is potentially a major tool for objective textbook evaluation, both for revision and for selection decisions. Thabet (1984) describes such a study of textbooks in use in Egypt during the 1970's (Living English, Progressive Living English and Practice and Progress) using TXTPRO under SAS (see Tang, 1985).

2. Analysis of large text databases

If genuine rather than constructed examples are to be used in teaching, it is necessary to have access to genuine text. Until quite recently, there was difficulty in getting large quantities of text to process. As Sinclair (1982) points out, this problem

is fast disappearing. The spread of word processing means that text is becoming routinely available in computer-readable form. Therefore, the problem is becoming how to select the right texts for analysis.

Storage media are no longer a constraint, mainly through development of optical media. For example, Steven Jobs' NeXT computer has an optical disk drive, and comes equipped with the complete works of Shakespeare. The development of the compact disk as a storage medium (CD-ROM), is making large commercial texts and databases available in computer-readable form. Much of the material presently obtainable is specialist in nature - statistical tables, large text databases (such as ERIC), medical and scientific texts; however, there are texts of interest to ESL/EFL specialists. These include Grolier's Electronic Encyclopedia, News Digest, and The CIA World Fact Book (any CD-ROM supplier will be able to provide an up-to-date list). A concordancer can be used on these large texts to provide examples of usage, textual material for testing, and information about register (e.g. for ESP course design). The technology required is already within the price range of a teachers' center or curriculum development group, and likely to fall; therefore, this approach is likely to become much more general in the future.

3. Local or individual applications

Not all corpora have to be large research or commercial accumulations. A corpus can consist of any appropriate text, if the analysis has a specific, limited purpose. The same software used for major academic applications can be used for these local materials. There are numerous ways in which this empirical perspective could benefit ELT. Examples include:

- a. Classroom Applications:** in addition to the collection of examples for teaching, Windeatt (1987) gives four ways in which concordancing might be used in the language classroom - (i) to construct gapped exercises; (ii) to analyze learners' work and compare it with authentic examples; (iii) for error analysis; (iv) for enabling learners to carry out their own concordance searches. Concorde (see section E below) was developed at Sultan Qaboos University for this purpose. This approach is mainly applied in advanced academic English courses.
- b. Worksheets or materials:** The procedure suggested for textbook evaluation above can also be used to check content and appropriacy of examples in worksheets and other materials, including tests. Where a particular word or structure is practiced, the author can check the context for genuineness and importance; and hence determine empirically whether it is being used correctly, and whether supplementary examples of other usages might be justified.

- c. **Teacher's books:** Teacher's books are too often neglected. Many third-world teachers do not have a high standard of English, or much exposure modern methodology. The teacher's book is a vital tool in helping these teachers to update themselves and teach effectively. However, they may find teacher's books very difficult to read, because the lexical items used in describing teaching are not normally taught in school, and teacher's books written for native speakers may contain some quite difficult structures. A concordancer can be used to list the lexis and structures used, with frequencies and contexts. This information can be used for evaluation, and for revision by suggesting simpler (because well-known) synonyms, and less difficult structural patterns. When the text is revised, substitutions can be made automatically by a word processor, and the contexts checked. The information gathered from analysis of teacher's books can also be the basis of a training module (for both pre- and in-service use) based on the actual language the teacher will encounter in the teacher's books they have to read.
- d. **Transcripts of classroom tapes:** another essential input into a teacher training course is the language of classroom management. A concordancer can be used to classify the classroom management language of typical teachers using transcribed audiotapes of "good" classrooms. The product of this transcription and analysis can be used as the basis for both training and assessment materials, of the form "how do you handle this situation"; and "what do you say if ..."; and for producing error-correction tests and exercises.
- e. **Student writing samples:** entry of student writing samples to the computer makes possible repeated analysis. This is particularly attractive for the analysis of large amounts of material; the amount that can be processed is effectively limited only by the funds available for data entry. A concordancer can be used to check the range of words and structures used/not used; whether the words or patterns produced show accuracy of understanding; the presence of specified interlanguage features. It can also be used to check one student against another, or against the group as a whole, on any and every feature.
- f. **Academic writing:** where a supply of academic writing is available (e.g. by using faculty members' own word-processed production, or where the text of papers accepted by a journal is available), this can be used in two main ways. First, it could be the basis of a study of the register - e.g., Egyptian academic writing - and the behavior of academics - citation behavior, for example. This could be the foundation for the writing components in an EAP course.

Second, it could be used to give guidance to those whose work is included, seeking to master academic writing, or submitting their work to a journal.

E. SELECTING A SUITABLE SOFTWARE PROGRAM

Because the early concordance programs were written for mainframe computers, they were neither easily accessible nor very user-friendly. An example of such a program is TXTPRO (Tang, 1985). A microcomputer version of TXTPRO is now available, with valuable features; but it is difficult for inexperienced users to handle. Coker and Underwood (1981) describe a number of available programs, in the context of linguistic analysis of samples from bilingual children. More recently, programs especially designed for small computers have become available. Davison (1983) and Johns (1986) describe such programs.

The basic function of a concordance program is to provide a concordance; a set of contexts for each word. However, a lot of the value of this concordance depends on the precision with which the researcher can specify target items and their presentation. This in its turn depends in part on additional features, including the wordlist and index. The following are the main factors to be considered in selecting a program:

1. **Range of facilities supported.** Any program will do a complete concordance (though there may be limitations on length of text). Most concordancers also allow some specification of which individual lexical items should be processed. Classroom use will usually require only the basic "concordancing" function. However, a researcher's decisions about strategy will depend on the distribution of words and structures in the text. In this case, the concordance can only be planned using a wordlist and index. Therefore, a concordancer that prepares a wordlist and index will save time spent on pre-processing the data using another program. This utility is not often found in cheaper concordance programs.

It is also valuable if the program can group items under "headwords" (e.g., different grammatical forms derived from a single root). This is a further refinement not usually found in low-cost programs.

2. **"wildcard" facility.** The ability to locate specific features during concordancing depends on the "wild cards" available within the program - for example, an "at" (@) commonly means "any one letter"; an asterisk (*) means "any number of letters including none". Wild cards enable the researcher to locate categories of items. For example, "*ing" would locate all words ending in "-ing". It could be used to find gerunds, such as "coming" and "going", but it would also throw up "thing" and

"bring", and these would have to be edited out. By contrast, "@ing" would locate only four-letter words ending in "-ing", such as "sing" and "ring". In good programs, wild cards can be combined to search for a phrase pattern. For example, "am *ing * to" would locate "... am looking forward to ..." and "... am going back to ..." but not "I am hoping to ...". With skills, it is possible to use wild cards to make a contextualized distribution for a specific morpheme, a single word, a phrase or grammatical pattern, as required.

3. **flexibility in selecting texts.** A corpus may consist of texts tagged in various ways. A corpus of plays, for example, could be tagged by author, act/scene, and speaker; a body of student writing might be tagged by topic, student, gender, and test score. This can be done in several ways. A common one is called "cocoa", and uses angle brackets to enclose the tags. For example, <A 1> would mean "Act 1", and <C Caesar> would mean "character - Caesar". Using a tagging system of this kind, it is possible to restrict analysis to only part of the corpus - for example, the words of one speaker in a play, or the teacher in a lesson transcript, a single student, or one essay topic. Most full-featured programs have a tagging system based on references inserted in the text; cheaper programs, and programs that operate by building an index of the corpus, may have a more limited facility. However, some programs attempt to compensate for not being able to process tags - Micro-concord, for example, can accept any number of text files, including use of wild cards in file names; and thereby achieves considerable flexibility; but lacking internal markers in files, cannot easily analyze components of a single text file.

4. **sorting:** sorting contexts in a concordance groups together occurrences of a particular pattern of words. A "right" sort (by contexts to the right of the keys) groups the words following the keys (i.e., all the occurrences of "go away" will be grouped, and separated from the examples of "go towards"). A "left" sort (by contexts to the left of the keys) does the same with the words before the keys. A "backwards" sort sorts the context to the left of the keys, from right to left. This kind of sort is useful for grouping word endings - all the words ending in "-ing" will be grouped separately from all the words ending in "-ed". Good sorting is therefore highly important in a concordancer.

5. **flexibility of context.** For most linguistic applications, KWIC is used. It is useful to be able to vary the depth of context in terms of number of characters, either the amount that fills a screen line, or the amount that the printer can print (to allow for a wide carriage, or condensed print). It is also useful to be able to specify whether the context is to the left or right of the key, or both. For some other applications, those more concerned with the rhetorical context (i.e., making a study of imagery, or allusions to a particular topic) it may be better

to define context as a meaningful unit - i.e., a sentence or paragraph. A good program will be able to handle more than one type of context.

6. Dealing with word-processed text. Most concordancers require text in American Standard Code for Information Interchange (ASCII), which uses seven bits. Many word processors use 8-bit characters, the eighth (or "high") bit being used for information about text format. In compiling a corpus it is necessary to include texts prepared on many word processors. Some word processors (such as Word Perfect) can convert text into ASCII. Otherwise, a utility is necessary for stripping high bits from word-processed text; or the program must accept text from common word processors such as WordStar.

Choosing a program

There are two main types of program available; high-cost fully-featured programs, and low-cost ones with fewer features, which because of their low cost are an excellent starting point, and also useful for students. Below are four examples of currently-available programs. Two are fully-featured programs; two lower-cost ones:

Micro-OCP: (Stg. #165) The PC version of a mainframe program used by the Oxford lexicographic team. Designed for literary and linguistic use. Instructions are given in a command language which allows very flexible selection of texts, keys (including powerful wild cards, selection of collocates and definition of headwords), contexts and sort strategies. Micro-OCP offers a word list with basic statistics (frequency profile giving cumulative total, relative frequencies, vocabulary and word totals, and type/token ratio), index and concordance. It can work with other languages (e.g. Spanish) using variations on the Latin alphabet. Published by Oxford University Press, 1988

Wordcruncher: (US \$299) Another very good program. It depends on pre-indexing, and has very flexible retrieval facilities. It can work with other languages than English, including Arabic. It is particularly aimed at professionals such as lawyers accessing large bodies of text. Produced at Brigham Young University; marketed by Electronic Text Corporation, 5600 North University Avenue, Provo, Utah.

Micro-concord: written by Tim Johns, Birmingham University; to be published by Oxford University Press late in 1989 at an expected cost of under Stg. #100. A development of an earlier Spectrum program (Johns, 1986). The beta-test version shows very good functionality for a low cost program. It offers use of a ramdisk to speed up processing, choice of disk to use for text files, manual or batch selection of texts, KWIC or sentence context, and KWIC context width at 74, 90, or 130 columns. It has very good

wild cards, capable of selecting sentence patterns (e.g. "be *ed * the" worked), and flexible sorting (left-, right- or backward sorts, with prioritizing and sorting to a depth of five words).

The Longman Concordance Program: to be published late in 1989. This program is a product of the Longman lexicographic unit. This is an indexing concordancer (like Wordcruncher). It has a good user interface and works very rapidly. It sort flexibly; but text size is limited (i.e., up to a novel-length book, but not a large corpus of texts).

Kaye: the IBM/SEU concordancer. Product of a collaboration between the University of London Survey of English Usage and IBM at Winchester, England. This is another indexing concordancer (similar to Wordcruncher). It uses a split screen, providing a sentence context for one citation in the upper half of the screen, and KWIC concordance for eight contexts in the lower half. Not commercially available.

Concorde: written by Dave Poulton with an interface by Vance Stevens (both of Sultan Qaboos University, Oman); available as shareware. This simple program does a basic concordance. It allows only one wild card, and will not accept patterns of more than one word. The program relies on the DOS SORT command (which is less flexible than, e.g., the Micro-concord sort). It offers only KWIC context. However, it is a very useful introduction.

F. CONCLUSION

Computer-based text concordancing is a new tool for English language teachers and researchers. This paper has reviewed its development, and suggested some possibilities for future use. It is the authors' view that these and other applications will develop rapidly; and that in the process a new, more empirical approach to the preparation and evaluation of materials will become possible. If so, this will be an exciting development for researchers and curriculum developers, students and teachers.

BIBLIOGRAPHY

- Andersson, Erik and Ostman, Jan-Ola Computer Processing of Swedish Syntactic Data. Some Preliminaries and Tentative Results. ERIC Accession Number: ED279151. 1978.
- Atwell, E. and S. Elliott. Dealing with ill-formed English text. in Garside, Leech and Sampson, 1987.
- Benson, M. Collocations and Idioms. in: Ilson, 1985.
- Bergenholtz, H. and B. Schaefer (eds). Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora. Königstein: Scriptor, 1979.
- Booth, B. Text input and pre-processing: Dealing with the orthographic form of texts. in Garside, Leech and Sampson, 1987.
- Coker, Pamela L. and Underwood, Mark A. Automatic Linguistic Analysis. Los Alamitos, Ca: National Center for Bilingual Research, 1981. ERIC Accession Number: ED234639.
- Collins COBUILD. The Collins COBUILD English Language Dictionary. London, Collins, 1987.
- Davison, Ned J. An Interactive Concordance Program for the Small Computer. CALICO Journal, v1 n1 p24-26 June 1983.
- Francis, W.N. "Problems of assembling and computerizing large corpora". in Bergenholtz and Schaefer (1979), 110-23.
- Francis, W.N. and H. Kucera. Frequency analysis of English usage: lexicon and grammar. New York: Houghton Mifflin, 1982.
- Garside, R., G. Leech and G. Sampson, eds. The Computational Analysis of English. Harlow: Longman, 1987.
- Ilson, R. (ed.), Dictionaries, Lexicography and Language Learning ELT Documents #20, Oxford: Pergamon Press, 1985.
- Johansson, S (ed.) Computer corpora in English language research. Bergen: Norwegian Computing Centre for the Humanities 1982.
- Johns, Tim. Micro-Concord: A Language Learner's Research Tool. System, v14 n2 p151-62 1986.
- Kennedy, Graeme D. Expressing Temporal Frequency in Academic English. TESOL Quarterly, v21 n1 p69-86, March 1987.

Kennedy, Graeme D. Discovering How a Communicative Notion Is Expressed in English. Paper presented at the annual meeting of the Teachers of English to Speakers of Other Languages (19th, New York, NY, April 8-14, 1985). ERIC Accession Number: ED259560.

Khaled, Turki Bani "Linguistic Analysis as an Input to Course Design". In: Proceedings of 3rd Annual Linguistics Conference (Irbid, Jordan, April 1-3, 1984) ERIC Accession Number: ED271012.

Leech, G. General Introduction. in Garside, Leech and Sampson, 1987

Leech, G. Automatic Grammatical Analysis and its applications. in Leech and Candlin, 1986

Leech, G. and C. Candlin (eds.), Computers in English Language Teaching and Research Harlow: Longman, 1986.

Lofstrom, Jonas. Reports on Swedish Projects Relevant to the Study of Impromptu Speech: Sprakdata. In: Enkvist, Nils Erik, Ed. Impromptu Speech: A Symposium. Papers Contributed to a Symposium on Problems in the Linguistic Study of Impromptu Speech (Abo, Finland, November 20-22, 1981). ERIC Accession Number: ED277241, 1982.

Macdonald, Ross and Others. Improving Techniques in Teaching English for the Job. Rosslyn, Va. National Clearinghouse for Bilingual Education, 1982.

Mitton, Roger. A Collection of Computer-Readable Corpora of English Spelling Errors. Reading Teacher, v39 n9 p931-33 May 1986.

Pickett, G. D. "Defining Business English." Paper presented at the Annual Meeting of the International Association of Teachers of English as a Foreign Language (20th, Brighton, England, April 1-4, 1986). ERIC Accession Number: ED274155. 1986.

Sinclair, J. "Reflections on computer corpora in English language research". in Johansson, 1982.

Sinclair, J. "Lexicographic Evidence." in Ilson, 1985.

Sinclair, J. "Basic Computer Processing of Long Texts". in Leech and Candlin, 1986.

Sinclair, J. "Introduction to the COBUILD Dictionary". London: Collins, 1987.

Svartvik, Jan and Others. Survey of Spoken English: Report on Research 1975-81. Lund Studies in English, No. 63. ERIC accession number: ED222038. 1982

Tang, Michael S. Microcomputers, Software and Foreign Languages for Special Purposes: An Analysis of TXTPRO. In: Proceedings of the Eastern Michigan University Conference on Languages for Business and the Professions (4th, Dearborn, MI, May 2-4, 1985) ERIC Accession Number: ED271986

Thabet, Ahmed Anwar Cohesion in EFL programs - a computational linguistics approach Georgetown University, Diss., 1983

Windeatt, S. Concordances in Language Teaching RECALL Computers in English Language Teaching and Research. No. 3 February 1987

Willis, Dave and Jane Willis. The Collins COBUILD English Course. London: Collins, 1987.